Freshwater Biology (2003) **48**, 1898–1911

APPLIED ISSUES

Measuring and controlling data quality in biological assemblage surveys with special reference to stream benthic macroinvertebrates

YONG CAO, CHARLES P. HAWKINS AND MARK R. VINSON
Department of Aquatic, Watershed and Earth Resources and Ecology Center, Utah State University, Logan, UT, U.S.A.

SUMMARY

1. Biological assemblage surveys primarily aim to characterise species composition and relative abundance at one or more spatial or temporal scales. Data interpretation and conclusions are dependent on how well samples characterise the assemblage of interest.
2. Conventional measures of data quality, e.g. standard deviations or coefficients of variation, were designed for single variable estimation, and they are either insufficient or invalid for assessing the quality of data describing entire assemblages. Similarity indices take species composition and relative abundance into account and may be used to effectively measure and control the quality of data used to characterise assemblage structure.
3. The average Jaccard coefficient (JC) calculated across multiple pairs of replicate samples, i.e. autosimilarity JC (AJC), is conceptually and numerically related to the average coefficient of variation in the densities of all species recorded, a measure of sampling precision, and to the proportion of total species richness sampled, a measure of sampling accuracy.
4. We explored how AJC can be used to assess the effect of different potential sources of error on the quality of assemblage survey data, including the sampling effort used both within regions and at individual sites, the individuals collecting samples, sub-sampling procedures, and consistency of taxon identification.
5. We found that the autosimilarity-based approach overcomes most weaknesses associated with conventional measures of data quality and can be used to effectively measure and control the quality of assemblage survey data.

Keywords: autosimilarity, autosimilarity Jaccard coefficient, bioassessment, community ecology, data quality, field surveys, macroinvertebrate assemblages, sampling design

Introduction

Biological assemblage surveys provide information on species composition and their relative abundance at various spatial and temporal scales. The information generated is used to address a variety of questions in ecology, natural resources and environmental management, including (i) determining spatial patterns of species diversity and species composition in relation to environmental factors (Gauch, 1982; ter Braak, 1987), (ii) testing ecological hypotheses and theories (MacArthur, 1965; Connell, 1978; Ricklefs & Schluter, 1993; Statzner, Resh & Roux, 1994) and (iii) detecting and quantifying the effects of human activities on ecosystem health (Rosenberg & Resh, 1993; Hawkins *et al.*, 2000; Wright, Sutcliffe & Furse, 2000).

Broad-scale surveys and long-term monitoring programs have greatly increased over the past two

Correspondence: Yong Cao, Department of Aquatic, Watershed and Earth Resources, Utah State University, Logan, UT 84322-5210, U.S.A. E-mail: yongcao@cc.usu.edu

decades. Sampling designs used in these programs often vary considerably in one or more of the following factors: the number and distribution of sampling sites, the sampling effort used at each site, the type of habitats sampled, the sampling method used, the frequency of sampling, the person collecting samples, the taxonomic resolution used during specimen identification and the person identifying specimens. A variety of errors and biases can be introduced into data at every step of their collection and analysis. The increasing need to compare and integrate both raw data as well as the results derived from different studies and monitoring programs requires that we specify the degree to which data collected in different ways are of comparable quality [Intergovernmental Task Force on Monitoring Water Quality (ITFM), 1995; Carter & Resh, 2001; Houston *et al*., 2002]. Measuring and controlling data quality has therefore become increasingly important for both basic ecological studies and bioassessments (e.g. Dines & Murray-Bligh, 2000; Humphrey, Storey & Thurtell, 2000; Mucina, Schaminée & Rodwell, 2000; Clarke *et al*., 2002).

The concept of data quality can sometimes be ambiguous because its definition depends on both the context and goal of a particular study. However, in a statistical sense, data quality consists of two elements: *accuracy* and *precision*. Accuracy means how close the observed value is to the true value, whereas precision measures how close repeated measurements are (Sokal & Rohlf, 1987; Zar, 1999). In assemblage surveys, the true values of many variables are not available, because obtaining them may require a census over the whole study area, and what is actually meant by accuracy has not been well defined. As a result, precision is often the only measure of data quality used and has been typically evaluated with the standard deviation (SD) or the coefficient of variation ($CV = SD/\bar{X}$) of the variable of interest. Both the SD and CV have been commonly used to measure the precision of estimating individual assemblage attributes and biotic indices (e.g. Turner & Trexler, 1997; Rabeni, Wang & Sarver, 1999; Clarke *et al*., 2002; Houston *et al*., 2002). Diamond, Barbour & Stribling (1996) provided a detailed overview about how the CV could be applied to data quality control in aquatic bioassessment.

Many factors that affect the quality of data in assemblage surveys have been addressed, including the choice of sampling sites within a study region (Stevens, 1994; Stevens & Olsen, 1999), sampling effort used at the local scale (e.g. Furse *et al*., 1981; Mackey, Cooling & Berrie, 1984; Angermeier & Smogor, 1995), who collected the samples (Mackey *et al*., 1984; Humphrey *et al*., 2000; Clarke *et al*., 2002) and laboratory procedures followed when processing samples (Doberstein, Karr & Conquest, 1999; Dines & Murray-Bligh, 2000). However, two major challenges remain. First, the effects of the different factors on data quality have not been examined systematically based on a common measure of data quality and therefore their relative importance can not be evaluated. For example, how species misidentifications, the number of sites sampled in a region and the sampling effort used at each site affect data quality has not been assessed on the same basis. Second, commonly used measures of data quality, such as the SD and CV, may be insufficient or irrelevant when assessing assemblage data quality for at least three reasons:

1. Assemblage surveys are usually designed to characterise species composition and relative abundance at a spatial scale of interest (e.g. stream reach, catchment or ecoregion). However, no single assemblage attribute can represent how well we have measured overall assemblage structure. For example, similar species richness can occur across a set of samples that vary considerably in species composition, and similar densities can occur across samples that differ greatly in both species richness and composition. The difference between characterising multivariate systems and estimating a single variable has been largely ignored, as pointed out by Kenkel, Juhasz-Nagy & Podani (1989), and was only addressed recently by Cao, Williams & Larsen (2002a). Assemblage data are often analysed with multivariate approaches, such as ordination and cluster analysis (ter Braak, 1987; Legendre & Legendre, 1998), but precision estimated for individual attributes or indices does not provide a basis to assess the quality of assemblage data used in multivariate analysis. For example, there is no necessary relationship between high precision in species richness estimates and either the stability or statistical significance of outcomes derived from multivariate analysis.

2. When assemblage data are collected for estimating different biotic indices and other assemblage attributes, the variability in those measures often differs (Resh & McElravy, 1993; Doberstein *et al*., 1999;

Sovell & Vondracek, 1999). Thus, data quality needs to be controlled individually for each index or attribute, but doing so is not possible because these measures are normally derived from the same raw data. Consequently, the precision of different indices and assemblage attributes will differ under the same data-quality procedure.

3. Estimates of many assemblage attributes and biotic indices are dependent on sampling effort. We refer to 'sampling effort' as the sampling area or volume or the number of individuals counted. As the size of each sampling unit is standardised, sampling effort can also be expressed as the number of sampling units pooled. A sampling unit may be a plot within a site or a site within a region. Measures whose values can change with sampling effort include species richness (Colwell & Coddington, 1994; He, Legendre & Bellehumeur, 1994), diversity indices (Hughes, 1978; Pinder *et al.*, 1987) and biotic indices (Pinder *et al.*, 1987; Stark, 1993). Precision of individual assemblage attributes, diversity indices and biotic indices can also be dependent on sampling effort, which means that some measures of precision will be sampling-effort specific.

A recently developed concept, *autosimilarity* (Cao *et al.*, 2002a), may provide a way of measuring and controlling the quality of assemblage survey data at all major steps of data collection and analysis. Autosimilarity is the average compositional similarity observed across multiple pairs of replicate samples (Cao *et al.*, 2002a). We define the term replicate sample as a single sample unit (e.g. a quadrat or plot), or a given number of sample units pooled, or a fixed number of individuals counted. Autosimilarity can be measured with various similarity indices, although the Jaccard coefficient (JC) has especially useful properties. The level of autosimilarity indicates how well a set of sample units pooled characterises the whole assemblage.

In this paper, we (i) demonstrate how particular measures of autosimilarity are explicitly related to both accuracy and precision for multivariate systems in both conceptual and numerical terms, (ii) illustrate how autosimilarity measured with the Jaccard coefficient (AJC) can be used for assessing and controlling the effects of all major sources of error in assemblage surveys with special reference to aquatic bioassessment and (iii) discuss the practical implications and limitations of this data-quality control approach.

Using autosimilarity for data quality control: conceptual and numerical justifications and its estimation

Conceptual considerations

An assemblage survey usually yields a species by site matrix. From this matrix, a variety of biotic and diversity indices can be derived and multivariate analysis can also be conducted. The samples used to populate this matrix should therefore *accurately and precisely represent both species composition and their relative abundances*. In this context, *precision* refers to the similarity in species composition and relative abundance between any two replicate samples, which is exactly what autosimilarity measures. Later we will show that AJC is also a numerical extension of the coefficient of variation in multivariate systems.

The most accurate sample is expected to capture every species in the sampled assemblage and describe the relative abundance of each species. Species relative abundance in an assemblage usually stabilises with less sampling effort than is required to capture all species (Angermeier & Smogor, 1995; Cao, Williams & Bark, 1997). Therefore, the percentage of the total species richness sampled (%TSR) is a simple measure of accuracy. For example, a sample capturing 95% TSR at a site more accurately represents the local assemblage than a sample capturing 20% TSR. Furthermore, estimates of %TSR are not based on a random subset of the total species richness (TSR) in an assemblage. Instead, common species are always collected first. Therefore, X% TSR basically means the most common X% of species in the assemblage have been captured. The difficulty is that TSR is usually unknown. Although numerous statistical methods have been proposed to estimate TSR from sampling data (Colwell & Coddington, 1994; Williams, Nichols & Conroy, 2001), it remains a great challenge to estimate TSR accurately and precisely based on a small set of samples (Longino, Coddington & Colwell, 2002; Wagner & Wildi, 2002; Petersen & Meier, 2003). However, previous studies have demonstrated that %TSR is strongly and consistently related to AJC across a variety of assemblages (Cao, Larsen & Hughes, 2001b; Cao *et al.*, 2002a). This relationship is quite easy to understand. Assuming that the sampling method used does not selectively miss any species, 100% TSR is reached when all replicate samples share every species. In contrast, when few species are common among

replicate samples, a low %TSR is expected. Estimates of the %TSR–AJC relationship from samples is also stable unless TSR is very low (Y. Cao, unpubl. data). As a result, both sampling accuracy and precision can be measured with a single statistic, autosimilarity (AJC).

Among numerous similarity indices available (Legendre & Legendre, 1998), the JC is especially useful for measuring autosimilarity because of its simplicity, statistical explicitness, consistent correlation with %TSR (Cao *et al.*, 2002a) and explicit linkage with the average CV of species abundance across all species (see below). Other binary indices, such as the Sørensen Index, might also be of potential use, but their relationships with %TSR are uncertain and they can not be clearly related to the average CV. When relative abundance is considered important, abundance-based similarity indices, such as the Bray-Curtis Index, may be applied. However, their relationships with %TSR can be complex and inconsistent across assemblages. The choice of data transformation, e.g. log (x) versus $\sqrt{x}$, can further complicate the interpretation of autosimilarity values. As a result, at this time we focus only on the use of the JC.

Numerical considerations

Here, we demonstrate how AJC can be numerically related to the mean CV of species abundance across all species. Assuming that only two replicates are collected from a site containing n species, we have a replicates × species matrix as below:

Species	Replicate 1	Replicate 2
1	X_{11}	X_{21}
2	X_{21}	X_{22}
·	·	·
n	X_{1n}	X_{2n}

X_{ij} is the number of individuals of species j in replicate i. For species j, we can calculate the SD_j and CV_j across the two replicates:

Given $\overline{X_j} = \frac{X_{1j}+X_{2j}}{2}$, then

$$SD_j = \sqrt{(X_{1j} - \bar{X}_j)^2 + (X_{2j} - \bar{X}_j)^2} = |X_{1j} - X_{2j}| \times \sqrt{0.5} \quad (1)$$

and

$$CV_j = \frac{SD_j}{\bar{X}_j} = \frac{\sqrt{0.5}}{0.5}\left(\frac{|X_{1j} - X_{2j}|}{X_{1j} + X_{2j}}\right) \quad (2)$$

An average CV across all n species gives an overall evaluation of precision. We then have:

$$\overline{CV} = \frac{1}{n}\sum CV_j = \frac{\sqrt{2}}{n}\sum \frac{|X_{1j} - X_{2j}|}{X_{1j} + X_{2j}} \quad (3)$$

When we consider species presence–absence only:
if c species are common to both replicates,

$$\sum_{j=1}^{c} \frac{|X_{1j} - X_{2j}|}{X_{1j} + X_{2j}} = 0$$

if a species are present in replicate 1 only,

$$\sum_{j=1}^{a} \frac{|X_{1j} - X_{2j}|}{X_{1j} + X_{2j}} = a$$

if b species are present in replicate 2 only,

$$\sum_{j=1}^{b} \frac{|X_{1j} - X_{2j}|}{X_{1j} + X_{2j}} = b$$

Given that $a + b + c = n$, we have:

$$\overline{CV} = \frac{\sqrt{2}}{n}(a+b) = \sqrt{2}\left(1 - \frac{c}{a+b+c}\right) \quad (4)$$

Therefore,

$$\overline{CV} = (1 - JC) \times C \quad \text{where } C = \sqrt{2} \quad (5)$$

As C is a constant, we divide this CV by C to make it range between 0 (least precision) and 1 (most precision). CV/C is the dissimilarity form of JC.

Estimating AJC from samples

When N sample units or individuals are collected, they can be randomly divided into m distinct groups with n units per group ($m = N/n$). Each group is pooled to create a composite sample. AJC is obtained by averaging JC for all possible comparisons of these composite samples. However, with an increase in n, the number of distinct groups (m) rapidly decreases. At 1/2 N-units of sampling effort, only two distinct samples can be created. Because the JC value can be dependent on how the N sample units are divided, a re-sampling procedure can be used to overcome this difficulty by repeatedly and randomly dividing the N sample units into two groups. AJC is then obtained by averaging the values of the JC from

these different random combinations (Cao *et al.*, 2002a).

Because of heterogeneity in species composition across samples, the estimation of AJC may not be precise when the number of sample units or the number of individuals is too small. A minimum number of sample units (e.g. >10) or individuals (e.g. 300) should be used to allow estimation of AJC from a large number of combinations, but the exact number will depend on the level of AJC desired. We address this issue further later in the paper.

Using AJC to measure and control data quality in assemblage surveys

Precision and accuracy take different forms at different steps of data collection or for different sources of error. We will therefore independently establish the relevant criterion of data quality and describe how AJC can be used to measure and control data quality for each major data collection step and source of error. Where possible, we used real data to illustrate the procedures.

Characterising regional assemblages

Regional surveys are used to determine the overall condition of biological resources or the distribution of site conditions within a large area. Two major questions that arise when designing a survey are: how many sites should be sampled and what specific sites should be selected? Both design considerations should be addressed in such a way that the sampled sites adequately characterise both species composition and the distribution of species within the region of interest. Two types of survey designs have been commonly used for choosing sampling sites (Stevens, 1994). An empirical design targets sites of different types in proportion to their occurrences in the environment. In a statistical design, every element in a 'population', such as all streams or lakes in a region, has some chance to be sampled; and site selection is carried out by a randomisation procedure (Stevens, 1994; Hughes, Paulsen & Stoddard, 2000). Both designs have their advantages and disadvantages, however neither answers the question of how many sites should be sampled. At the regional scale, each sampling site can be regarded as a sample unit. When N sites are selected and sampled, based on either the empirical or statistical designs, AJC can be calculated following the re-sampling procedure described earlier. The level of AJC indicates how well the biota in the region is characterised by the N sites. If AJC = 1.0, all species present in the region were collected at the sample sites, and those sites thus represent the whole region with 100% precision and accuracy. The actual level of AJC targeted will be <1 because of resource constraints. Because large-scale surveys often cover large regions (e.g. Wright, 1995; Hawkins *et al.*, 2000; Reynoldson, Rosenberg & Resh, 2001), the desired level of AJC may not be reached with 1 year of sampling. In such cases, more sampling sites can be added in subsequent years until a target level of AJC is reached.

A regional survey can be further improved by applying the same AJC standard to sub-regions, e.g. ecoregions (Omernik, 1995; Omernik & Bailey, 1997) or catchments. Such standardisation would improve data comparability among the sub-regions and use research resources more effectively (i.e. by avoiding over-sampling in one sub-region and under-sampling in another). Consider two sub-regions A and B. In region A, species composition varies more substantially across sites than in region B. The former logically will require more sites to be characterised than the latter. Below, we use a real data set to illustrate how this procedure can be applied to measure and control data quality in the regional surveys.

More than 2000 macroinvertebrate samples were collected from North Carolina streams between 1983 and 1992 (Lenat, 1993). We used a subset of 209 samples from reference sites to determine how well these samples represented reference conditions in the three ecoregions of the state: Coastal Plain (33 samples), Mountain (142 samples) and Piedmont (34 samples). We estimated AJC at all possible sampling levels for each of the three regions following the procedure described earlier (Fig. 1). The Mountain region reached the highest AJC (0.72). Piedmont and Coastal Plain ecoregions were represented less well, with AJC values of 0.63 and 0.57, respectively. These differences could have resulted from either different numbers of sites in the three ecoregions or differences in ecological heterogeneity within each ecoregion or both. We therefore compared the three ecoregions based on the same number of samples. With $N = 32$, the Coastal Plain reached a slightly lower AJC (0.56)

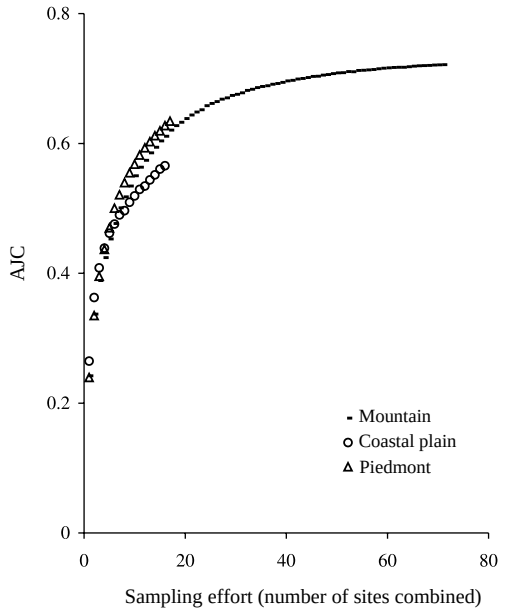

Fig. 1 Autosimilarity measured with Jaccard coefficient (AJC) increased with sampling effort (number of sites combined) in three ecoregions of North Carolina, but reached different AJC levels for the same sampling effort, indicating that these ecoregions were not characterised equally well by the sampling data.

than the Mountain (0.61) and Piedmont (0.63) ecoregions. The observed differences in AJC values among the three ecoregions appeared to result mainly from the different numbers of sites sampled. However, the lower value of AJC for the Coastal Plain implies that streams in that region may be more ecologically heterogeneous than streams in the other two regions and would therefore imply that this ecoregion would require more samples to adequately characterise it.

Characterising sites and comparing samples across sites

Site sampling should accurately and precisely characterise species composition and relative abundance within local assemblages and do so equally well across all sites to ensure data comparability. In other words, data quality associated with site sampling depends on two factors: how to standardise sampling effort and at what level. A standard sampling effort (e.g. area sampled or the number of individuals counted) often yields different AJC values and %TSRs at different sites (Cao *et al*., 2002a). This inequality of sampling accuracy and precision across sites will often result in underestimates of the true difference in richness and composition among sites because more diverse assemblages will be underrepresented by a standard sample relative to less diverse assemblages. For example, a Surber sample might capture 80% of a total of 20 taxa (16) at one site, but only 30% of a total of 60 taxa (18) at a more species-rich site, at which most of the taxa are typically rare. The observed richness ratio between these two sites is 16/18 or 8/9, whereas the true ratio is 20/60 or 1/3. Standardising sampling effort on AJC, rather than sampling area or the number of individuals, can overcome this difficulty (Cao *et al*., 2002a). It follows that the level of AJC on which sampling effort is standardised can significantly affect our ability to biologically distinguish sites from one another. Use of a high AJC level for standardisation means that most species are included in comparisons, which should therefore result in clearer and more accurate descriptions of the true biological differences among sites (Cao *et al*., 2002b). Below we used a macroinvertebrate data set to illustrate the procedure of standardising sampling effort on AJC in site sampling and to evaluate the AJC levels reached by commonly used sampling efforts.

Two stream sites in Idaho (ID-1 and ID-2) were extensively sampled in 2000 by six different field crews (Henderson *et al*., 2000). At each site, four riffle units were delimited and two Surber samples were collected from each riffle for a total of eight Surber units from each site. Samples were pooled in the field and all macroinvertebrates in the sample were later identified to the lowest possible level. Eighty-one taxa were recorded from all six sets of samples (48 Surber samples in total) at ID-1, and 97 taxa were collected at ID-2. Use of the Jackknife-2 method (Colwell, 1998) to estimate total taxa richness (TTR) resulted in estimates of 95 taxa at ID-1 and 120 taxa at ID-2. Jackknife-2 has been reported to work well at high sampling efforts (Palmer, 1990; Baltanás, 1992).

Using the re-sampling procedure described earlier, we calculated AJC at different fixed counts for each site in the two data sets. This analysis led to the following observations. First, the same sampling effort resulted in different values of AJC at different sites (Fig. 2a), indicating that different fixed counts would be required to obtain the same AJC level. For example, an AJC of 0.65 required 200 counts on average at ID-1, but about 800 counts at ID-2. Second, %TTR was strongly and consistently correlated with AJC at both sites (Fig. 2b), implying that the same AJC corresponded to similar %TSRs. Third, 100-count subsampling commonly used in bioassessment programs (Carter &

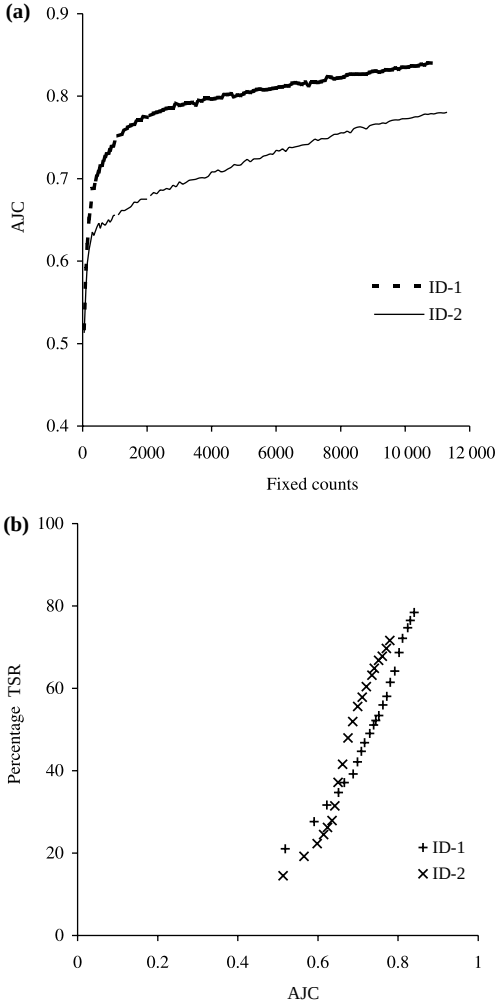

Fig. 2 Autosimilarity measured with Jaccard coefficient (AJC) increased with sampling effort (fixed-count) at two Idaho stream sites (a) and AJC is strongly and consistently correlated with %TSR (b).

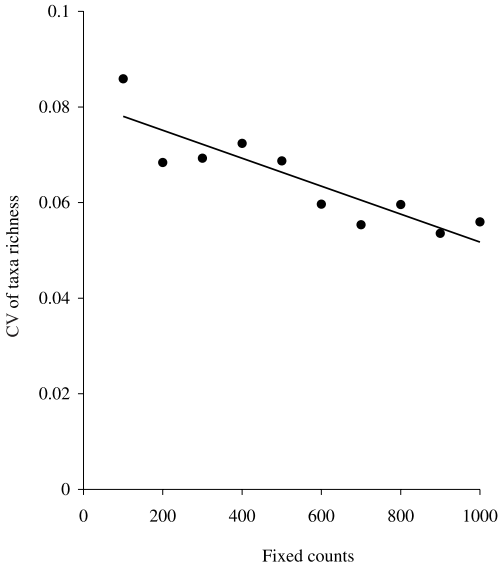

Fig. 3 The coefficient of variation (CV) of taxa richness per fixed-count sample decreased with increasing sampling effort at one Idaho stream site (ID-1).

Resh, 2001) resulted in AJC values of 0.56 and 0.59 and %TTR estimates of 19 and 28%. 500-count subsamples resulted in AJC values of 0.65 and 0.71 and %TTR values of 37 and 45%.

We also examined the effect of sampling effort on the CV of taxa richness. All individuals from ID-1 were randomly divided into multiple, distinct fixed-count subsamples at 10 levels of sampling effort (100–1000 counts at intervals of 100). Mean taxa richness per subsample and standard deviations were calculated. The CV significantly decreased with increasing fixed counts (Fig. 3) implying that when giving a CV value for taxa richness, one needs to specify the sampling effort used (e.g. CV = 0.15 for 100 counts). The CV value alone is difficult to interpret. For example, is a CV value of 0.2 in taxa richness precise? It may be, if it is based on 100 counts, but it might not be if based on 500 counts. Even with known sampling effort, it remains difficult to compare CV values from different studies. For example, does a CV of 0.10 in taxa richness based on 500 counts in one study imply higher precision than a CV of 0.15 based on 300 counts in another study? We do not know. In contrast, a high AJC (e.g. 0.9) always means high accuracy and high precision, and vice versa, regardless of the number of individuals counted or sample units pooled. This is another advantage of AJC over the CV in addition to its explicit linkage with %TSR.

As is the case for regional surveys, what AJC level is sufficient for inter-site comparisons will be dependent on the objectives of particular studies and the resources available. However, a low AJC level means a large proportion of the species pool was not sampled. Low levels of AJC may not affect the detection of strong ecological patterns or biological responses to environmental gradients, but they may prevent us from recognising weaker patterns or gradients or the initial stages of degradation or recovery. In addition, a low AJC would limit both the comparability of the data with other data sets and the general value of assemblage data in

addressing many general ecological questions, such as how species turnover rates and biodiversity patterns vary with landscapes and other environmental variables.

Effects of sample collectors

Long-term monitoring and large-scale surveys often involve multiple sample collectors. Technical training is usually given to ensure collectors follow the same operation procedures as much as possible (Barbour *et al*., 1999). However, it is important to know how much variation in data values is associated with inconsistencies among different collectors or the same collector over time.

This issue has been addressed in several studies for assemblage attributes and biotic indices. For example, Reynoldson *et al*. (2001) assessed the effect of different collectors on estimates of total number of individuals and taxa richness by conducting ANOVAs. Clarke (2000) reported that 12% of the total sampling standard deviation in taxon richness was attributed to collectors. Clarke *et al*. (2002) examined the effect of collectors on two biotic indices (the number of taxa in Biological Monitoring Working Party score system and Average Score Per Taxon) and reached similar conclusions.

The AJC can be used to assess the effect of collectors on species composition in samples. Inconsistencies in sampling may increase the variation in both species composition and relative abundance among replicates (i.e. decreasing precision). Consistent, but non-standard sampling can reduce accuracy. AJC calculated for multiple replicates collected by a single collector quantifies the precision of the collector. AJC calculated for a set of replicate samples collected by a group of collectors quantifies the precision across the group. However, in this case, accuracy is difficult to define because we do not know the true assemblage. In such a case, the sample collected by an experienced ecologist (trainer) may be used as a reference against which performance of others can be assessed. A variety of significance tests, including randomisation tests, could be applied here to test for differences among individuals.

We used the Idaho data set described earlier to illustrate how AJC can be used to assess the performance of sample collectors. As the eight Surber samples from each collector were pooled in the field, it was not possible to examine the precision and accuracy of collecting individual Surber samples. However, we conducted similar analyses on fixed-count subsamples taken from the pooled samples. The six collectors were treated equally, i.e. they were not identified as trainers or trainees. First, we defined a 'reference sample' and 'reference subsample'. Given that the complete taxon list was not available, we pooled all the samples except the one from the collector being evaluated to create a reference sample so that the reference was independent on the collector. Then, we randomly drew a fixed-count sample without replacement from the reference sample, taking it as a 'reference subsample'. A random fixed-count subsample was also drawn from the collector's sample. JC was calculated between these two subsamples. This process was repeated 1000 times to obtain AJC at each sampling level, which ranged from 100 to 2000 counts. In parallel, we established a taxon accumulation curve for each collector by randomly re-sampling his/her sample (1000 runs).

Differences in AJC among the six collectors were small, 0.04–0.08 at ID-1 and 0.05–0.07 at ID-2 (Fig. 4a,b), implying that the performances of the six collectors were generally consistent. The taxon accumulation curves showed a similar pattern (Fig. 4c,d), but there were some subtle differences. At ID-1, collectors 3 and 6 produced similar richness curves, whereas their AJC curves differed considerably. At ID-2, collector 3 collected fewer taxa than any other for the same fixed-count, but his AJC curve was similar. We also noted that the differences among collectors increased at higher sampling efforts, suggesting that at low sampling effort, the effect of collectors could be easily overridden by other error sources, such as random sampling error and the effect of heterogeneity.

Sub-sampling

The number of individuals in a macroinvertebrate sample is sometimes so large that collections must be sub-sampled. Ideally, a subsample should accurately represent the whole sample. Two factors influence data quality in this step: (i) subsamples may not be drawn at random and (ii) the subsample may be too small to characterise the whole sample. As we have addressed the second question earlier in this paper, we focused on the first question below.

In the Idaho data, approximately 500-count subsamples were manually drawn from each sample first

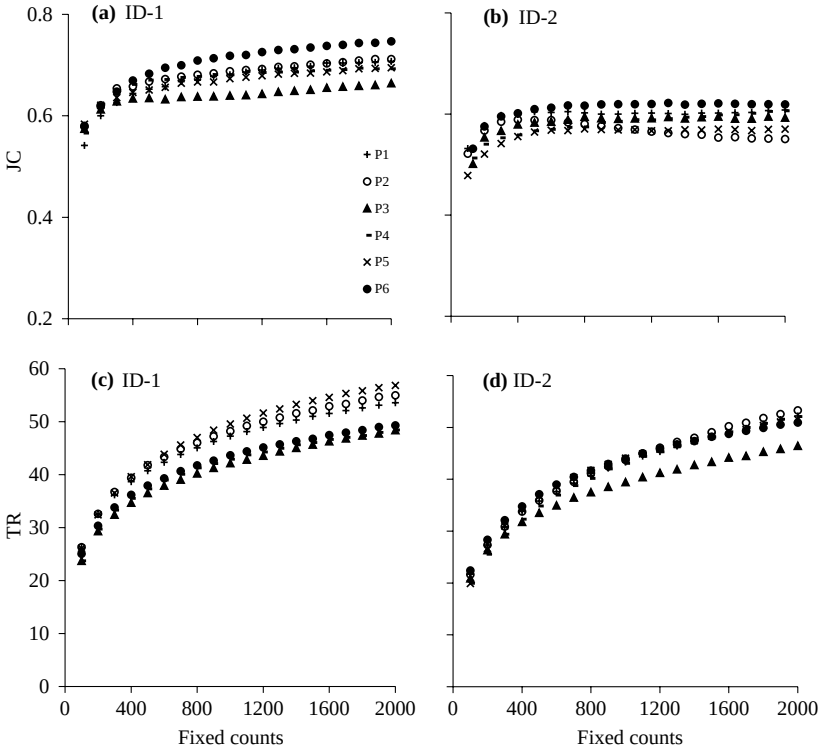

Fig. 4 The mean Jaccard coefficient (JC) between a random sub-sample from a collector and a random sub-sample from the reference sample varied among collectors and with increasing sampling effort at ID-1 (a) and ID-2 (b). Taxa richness (TR) showed a similar trend (c and d).

and then the rest of each sample was fully processed. We tested if the first approximately 500 individuals represented a random subsample as follows. Two thousand subsamples of the same size as the first subsample were randomly drawn from the whole sample. We then calculated JC between each of these subsamples and the whole sample. These 2000 JC values were then ordered from the lowest to the highest, and the 100th and 1900th values were taken as the lower and upper 95% confidence limits of JC estimates, respectively (see Pillar, 1998). The JC value between the first manually derived subsample and the whole sample was also calculated and was compared against the 95% confidence limits. The lower and upper 95% confidence limits were also determined for taxon richness in a similar way. There was no significant difference between the first subsamples and the random subsamples for either taxa richness or JC except for Collector Five at ID-1, whose subsample was significantly different, but only slightly lower, than the random subsamples in both measures (Table 1). This result indicates that errors associated with the laboratory subsampling procedure appeared to be random and did not affect data quality in any significant way.

Taxa identification and counting

Species can be misidentified and miscounted, which can affect data quality. A similarity index, such as the Jaccard coefficient, can be used for data-quality control at this final step of data collection. If two taxonomists, A and B, agree on all N species, JC = 1; if they disagree on n species, JC = 1 − (n/N). If taxonomist A acts as the reference, JC is the proportion of all species that were correctly identified by B. Multiple samples can be used to obtain an average JC for assessing the quality of species identifications by taxonomist B.

The U.K. Environment Agency implemented data-quality control for taxa identification and counting in their RIVPACS program (Dines & Murray-Bligh, 2000). The agency re-examined 60 samples each year from each regional laboratory and calculated the average missing taxa per sample. Two taxa were recommended as the maximum number allowed to be missed per sample. If 40 taxa are present in a sample, two missing taxa is equal to a JC of 0.95 (38/40), however if 20 taxa are present, the JC is equal to 0.90 (18/20). Thus, JC is a better measure of data quality for taxon identification than the absolute number of

Table 1 A test of random fixed-count subsampling for two Idaho sites: taxon richness and composition based on 2000 random runs (*5% significance level)

Crew	First fixed-count subsample	Taxa richness				Jaccard coefficient			
		Mean	95%-Upper	95%-Lower	Observed	Mean	95%-Upper	95%-Lower	Observed
ID-1									
1	486	40.37	44	36	39	0.7223	0.7857	0.6607	0.6964
2	522	42.12	46	38	43	0.6179	0.6765	0.5588	0.6613
3	570	37.52	41	34	39	0.7499	0.8200	0.6800	0.6954
4	509	37.57	41	34	35	0.6598	0.7193	0.5965	0.6753
5	557	43.10	47	39	38*	0.7561	0.8246	0.6842	0.6736*
6	446	37.86	41	34	35	0.7260	0.8039	0.6667	0.6755
ID-2									
1	473	35.27	39	31	36	0.5786	0.6393	0.5082	0.5902
2	429	34.48	39	30	38	0.6151	0.6786	0.5357	0.6325
3	428	32.26	36	29	32	0.5653	0.6316	0.4912	0.6092
4	466	33.94	38	30	35	0.5248	0.6000	0.4615	0.5900
5	437	35.23	39	31	36	0.5987	0.6780	0.5254	0.5781
6	447	36.03	40	32	35	0.6289	0.7018	0.5614	0.5620

taxa missed because it is not affected by the number of taxa in a sample. The Bray-Curtis index has been similarly used for data-quality control in taxa identification by Unicomarine Ltd (1996) for benthic invertebrates and Kelly (1999, 2001) for benthic diatoms.

The taxonomic resolution used can affect the estimation of AJC. Higher taxonomic resolutions, such as genus- or family-levels, yield higher similarities than species-level data (Guerold, 2000). Therefore, when similarity values are used to measure data quality the same level of taxonomic resolution should be employed across different datasets or sites compared.

Discussion

Ecological assemblages exhibit heterogeneity in both species composition and abundance at all spatial and temporal scales (Kolasa & Pickett, 1991; Palmer & White, 1994). This heterogeneity, together with the rarity of many species, makes it challenging to accurately or even consistently characterise assemblages when resources are scarce. Measures of data quality allow us to control or at least quantify estimates of sampling precision and accuracy. Such estimates provide a basis for establishing the confidence associated with the conclusions we draw and the interpretations we make. The key question is how to define and quantify data quality in a way that is both ecologically and statistically meaningful.

We described a framework that uses autosimilarity for measuring and controlling data quality and showed how it could be employed for assessing the effect of major sources of error in assemblage surveys. In general, we believe that this novel approach is advantageous over conventional measures of data quality, such as SD or CV, in a number of ways.

1. It focuses on the quality of species composition and relative abundance estimates (if an abundance-based similarity index is used), i.e. the raw data. Different data sets can thus be compared on the same basis.

2. AJC combines the concepts of both precision and accuracy together and measures assemblage data quality with a single interpretable, ecologically meaningful statistic.

3. Autosimilarity has a clearly defined range (0–1) and can be evaluated regardless of sampling efforts used and can be compared across regions, sites or data sets. In comparison, CV and SD have infinite ranges and they are specific for particular sampling efforts (Fig. 4), which makes the comparison of data quality across sampling effort or studies difficult.

4. Because both ordination and cluster analysis start with a similarity matrix (Green, 1980; Legendre & Legendre, 1998), autosimilarity is a statistically meaningful measure of data quality for multivariate analysis.

We have focused on illustrating how AJC can be applied to a few general types of situations. Use of AJC under specific conditions or to address specific questions will require careful consideration of the properties and limitations of AJC as a measure of data

quality. Of the several possible questions that may require additional attention, we discuss three below.

The majority of species in natural assemblages are rare, and therefore the AJC or %TSR reached is necessarily associated with the number of rare species captured. The role of rare species in community analysis and aquatic bioassessment is an active area of debate (e.g. Faith & Norris, 1989; Cao, Williams & Williams, 1998; Marchant, 1999; Cao, Larsen & Thorne, 2001a). Ecologists usually cannot and may not need to capture every species when studying factors that structure natural assemblages or when assessing if human activities affect assemblages. In bioassessment studies, assessments based on data with a low AJC, which excludes most rare species, would likely detect major effects of disturbances. However, the confidence associated with rejecting a no-impact hypothesis based on a small proportion of the species pool, i.e. a low AJC, may deserve serious consideration.

In many respects, it would be useful to adjust conventional sampling effort 'on the fly' while in the field so that all samples have similar values of AJC. For large-sized taxa, such as trees, birds and fish, species can be identified on site by trained crews, and AJC could thus be calculated while in the field to determine when sampling should stop. For benthic invertebrates, plankton and other small-sized taxonomic groups, decisions regarding when to stop sampling cannot be made in the field because the identification of most taxa is done in a laboratory. A possible solution is to over-sample, bring a large amount of sampled material to the laboratory, and then standardise sub-sampling of the materials on AJC. Laboratory processing would involve sequential identification of several subsamples and continue until a targeted AJC was reached. A simple randomisation program for calculating AJC can be run on a computer in the field or laboratory.

One further question is the relative importance of different potential sources of error in assemblage surveys. Consistently characterising the assemblage at regional and site-scales appears most prone to error because the cost to access and sample a site results in a relative small number of sites being sampled (low AJC). After samples are transported to the laboratory, the cost of sub-sampling is relatively low and a higher AJC should and can be reached than that produced by 100–200 count subsamples. For taxonomic identification, the highest AJC standard should be applied. Unfortunately, there is no method available to measure overall data quality, i.e. the combined effects of all different sources of error. Such a measure might be expressed as a product of AJC across three key factors: the number of sampling sites, sampling effort at each site and the accuracy of taxa identification. For example, if we reach an AJC of 0.95 for taxon identification, 0.75 at the site-scale and 0.50 at the regional-scale, the overall data quality would be $0.95 \times 0.75 \times 0.50 = 0.36$. Other factors, such as sample collectors and sub-sampling may further influence data quality. Future studies should quantify the relative importance of different survey steps and other sources of error on data quality. With strict data-quality control in assemblage surveys, ecologists will be able to obtain a clearer picture of ecological patterns and more accurate and precise bioassessments.

Acknowledgments

We are grateful to D.P. Larsen and Tetra-Tech colleagues for their constructive comments on the early versions of this manuscript, and to the North Carolina Division of Water Resources for permitting us to use their survey data. This research was funded by U.S. EPA – Science to Achieve Results (STAR) Program Grant no. R-82863701.

References

Angermeier P.L. & Smogor R.A. (1995) Estimating number of species and relative abundances in stream-fish communities: effects of sampling effort and discontinuous spatial distributions. *Canadian Journal of Fisheries and Aquatic Sciences*, **52**, 936–945.

Baltanás A. (1992) On the use of some methods for the estimation of species richness. *Oikos*, **65**, 484–492.

Barbour M.T., Gerritsen J., Snyder B.D. & Stribling J.B. (1999) *Revision to Rapid Bioassessment Protocols for Use in Streams and Rivers: Periphyton, Benthic Macroinvertebrates, and Fish. EPA 841-D-97002*. US Environmental Protection Agency, Washington, D.C.

ter Braak C.J.F. (1987) Ordination. In: *Data Analysis in Community and Landscape Ecology* (Eds R.H. Jongman, C.J.F. ter Braak & O.F.R. Tongeren van), pp. 91–169. Pudoc, Wageningen, the Netherlands.

Cao Y., Williams W.P. & Bark A.W. (1997) Effects of sample size (number of replicates) on similarity measures in river Aufwuchs community analysis. *Water Environment Research*, **69**, 107–114.

Cao Y., Williams D.D. & Williams N.E. (1998) How important are rare species in community ecology and bioassessment. *Limnology and Oceanography*, **43**, 1403–1409.

Cao Y., Larsen, D.P. & Thorne, R. St-J. (2001a) Rare species in multivariate analysis for bioassessment: some consideration. *Journal of the North American Benthological Society*, **20**, 144–153.

Cao Y., Larsen D.P. & Hughes R.M. (2001b) Evaluating sampling sufficiency in fish assemblage survey: a similarity-based approach. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 1782–1793.

Cao Y. Williams D.D. & Larsen D.P. (2002a) Comparison of ecological communities-the problem of sample representativeness. *Ecological Monographs*, **72**, 41–56.

Cao Y., Larsen D.P., Hughes R.M., Angermeier P. & Patton T. (2002b) Sampling efforts affect multivariate comparisons of stream assemblages. *Journal of the North American Benthological Society*, **21**, 701–714.

Carter J.L. & Resh V.H. (2001) After site selection and before data analysis: sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. *Journal of North American Benthological Society*, **20**, 658–682.

Clarke R. (2000) Uncertainty in estimates of biological quality based on RIVPACS. In: *Assessing the Biological Quality of Fresh Waters: RIVPACS and Other Techniques* (Eds J.F. Wight, D.W. Sutcliffe & M.T. Furse), pp. 25–39. Freshwater Biological Association, Ambleside, Cumbria, UK.

Clarke R.T., Furse M.T., Gunn R.J., Winder J.M., Wright J.F., Sutcliffe D.W. & Furse J.F. (2002) Sampling variation in macroinvertebrate data and implications for river quality indices. *Freshwater Biology*, **47**, 1735–1751.

Colwell R. (1998) *Statistical Estimation of Species Richness and Shared Species from Samples – a User Guide to EstimatorS 5*. Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, U.S.A.

Colwell R. & Coddington J.A. (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B*, **345**, 101–118.

Connell J.H. (1978) Diversity in tropical rain forests and coral reefs. *Science*, **199**, 1302–1310.

Diamond J.M., Barbour M.T. & Stribling J.B. (1996) Characterizing and comparing bioassessment methods and their results: a perspective. *Journal of North American Benthological Society*, **15**, 713–727.

Dines R.A. & Murray-Bligh J.A.D. (2000) Data assurance and RIVPACS. In: *Assessing the Biological Quality of Fresh Waters: RIVPACS and Other Techniques* (Eds J.F. Wight, D.W. Sutcliffe & M.T. Furse), pp. 71–78. Freshwater Biological Association, Ambleside, Cumbria, U.K.

Doberstein C.P., Karr J.R. & Conquest L.L. (1999) The effect of fixed-count subsampling on macroinvertebrate biomonitoring in small streams. *Freshwater Biology*, **44**, 355–366.

Faith D. P. & Norris R.H. (1989) Correlation of environmental variables with patterns of distribution and abundance of common and rare freshwater macroinvertebrates. *Biological Conservation*, **50**, 77–98.

Furse M.T., Wright, Sutcliffe, Furse J.F., Armitage P.D. & Moss D. (1981) An appraisal of pond-net samples for biological monitoring of lotic macroinvertebrates. *Water Research*, **15**, 679–689.

Gauch H.G. (1982) *Multivariate Analysis in Community Ecology*. Cambridge University Press, Cambridge.

Green R.H. (1980) Multivariate approaches in ecology: the assessment of ecological similarity. *Annual Review for Ecology and Systematics*, **11**, 1–14.

Guerold F. (2000) Influence of taxonomic determination level on several community indices. *Water Research*, **34**, 487–492.

Hawkins C.P., Norris R.H., Hogue J.N. & Feminella J.W. (2000) Development and evaluation of predictive models for measuring the biological integrity of streams. *Ecological Applications*, **10**, 1456–1477.

He F., Legendre P. & Bellehumeur C. (1994) Diversity pattern and spatial scales: a study of a tropical rain forest of Malaysia. *Environmental and Ecological Statistics*, **1**, 265–286.

Henderson R.C., Abbruzzese C.J., Mellison C., Bouwes B., Archer E.K. & Kershner J.L. (2000) *Effectiveness Monitoring Pilot Project for Streams and Riparian Areas within the Upper Columbia River Basin – Annual Summary Report 2000*. Fish Ecology Unit, 860 N 1200 E, Logan, UT 84321, U.S.A.

Houston L., Barbour M.T., Lenat D. & Penrose D. (2002) A multi-agency comparison of aquatic macroinvertebrate-based stream bioassessment methodologies. *Biological Indicators*, **1**, 279–292.

Hughes B.D. (1978) The influence of factors other than pollution on the value of Shannon's diversity index for benthic macroinvertebrates in streams. *Water Research*, **12**, 359–364.

Hughes R.M., Paulsen S.G. & Stoddard J.L. (2000) EMAP-surface water: a multiassemblage, probability survey of ecological integrity in the U.S.A. *Hydrobiologia*, **422/423**, 429–443.

Humphrey C.L., Storey A.W. & Thurtell L. (2000) AUSRIVAS: operator sample processing errors and temporal variability – implication for model sensitivity. In: *Assessing the Biological Quality of Fresh Waters:*

RIVPACS and Other Techniques (Eds J.F. Wight, D.W. Sutcliffe & M.T. Furse), pp. 144–163. Freshwater Biological Association, Ambleside, Cumbria, U.K.

ITFM (1995) *The Strategy for Improving Water-quality Monitoring in the United States – Final Report of the Intergovernmental Task Force on Monitoring Water Quality*. Office of Water Data Coordination, US Geological Survey, Reston, Virginia Open-File Report 95–742.

Kelly M.G. (1999) Progress towards quality assurance of benthic diatom and phytoplankton analyses in the UK. In: *Use of Algae for Monitoring Rivers III* (Eds J. Prygiel, B.A. Whitton & J. Bukowska), pp. 208-215. Agence de l'Eau Artois-Picardie, France.

Kelly M.G. (2001) Use of similarity measures for quality control of benthic diatom samples. *Water Research*, **35**, 2784–2788.

Kenkel N.C., Juhasz-Nagy P. & Podani J. (1989) On sampling procedures in population and community ecology. *Vegetatio*, **83**, 195–207.

Kolasa J. & Pickett S.T.A. (1991) *Ecological Heterogeneity*. Springer-Verlag, New York.

Legendre P. & Legendre L. (1998) *Numerical Ecology*, 2nd edn. Elsevier, New York.

Lenat D.R. (1993) A biotic index for the southeastern United States: derivation and list of tolerance values, with criteria for assessing water-quality ratings. *Journal of North American Benthological Society*, **12**, 279–290.

Longino J.T., Coddington J. & Colwell R.K. (2002) The ant fauna of a tropical rain forest: estimating species richness three different ways. *Ecology*, **83**, 689–702.

MacArthur R.H. (1965). Patterns of Species Diversity. *Biological Review*, **40**, 510–533.

Mackey A.P., Cooling D.A. & Berrie, A.D. (1984) An evaluation of sampling strategies for qualitative surveys of macroinvertebrates in rivers, using pond nets. *Journal of Applied Ecology*, **21**, 515–534.

Marchant R. (1999) How important are rare species in aquatic community ecology and bioassessment? A comment on the conclusion of Cao et al. 1998. *Limnology and Oceanography*, **44**, 1840–1841.

Mucina L., Schaminée J.H. & Rodwell J.S. (2000) Common data standards for recording relevés in field survey for vegetation classification. *Journal of Vegetation Science*, **11**, 769–772.

Omernik J.M. (1995) Ecoregions: a spatial framework for environmental management. In: *Biological Assessment and Criteria – Tools for Water Resources Planning and Decision Making* (Eds W.S. Davis & T.P. Simon), pp. 49–62. Lewis Publishers, Boca Raton, Florida.

Omernik S.J. & Bailey R.G. (1997) Distinguishing between watersheds and ecoregions. *Journal of American Water Resources Association*, **33**, 935–949.

Palmer M.W. (1990) The estimation of species richness by extrapolation. *Ecology*, **71**, 1195–1198.

Palmer M.W. & White P.S. (1994) On the existence of ecological communities. *Journal of Vegetation Science*, **5**, 279–282.

Petersen F.T. & Meier R. (2003) Testing species-richness estimation methods on single sample collection data using the Danish Diptera. *Biodiversity and Conservation*, **12**, 667–686.

Pillar D.V. (1998). Sampling sufficiency in ecological surveys. *Abstracta Botanica*, **22**, 37–48.

Pinder L.C.V., Ladle M., Gledhill T., Bass J.A. & Mathew A.M. (1987). Biological surveillance of water quality-1: a comparison of macroinvertebrate surveillance methods in relation to assessment of water quality in a chalk stream. *Archives of Hydrobiology*, **109**, 207–226.

Rabeni C.F., Wang N. & Sarver R.J. (1999) Evaluating adequacy of the representative stream reach used in invertebrate monitoring program. *Journal of North American Benthological Society*, **18**, 284–291.

Resh V.H. & McElravy E.P. (1993) Contemporary quantitative approaches to biomonitoring using benthic macroinvertebrates. In: *Freshwater Biomonitoring and Benthic Macroinvertebrates* (Eds D.M. Rosenberg & V.H. Reshh), pp. 159-194. Chapman and Hall, London.

Reynoldson T.B., Rosenberg D.M. & Resh V.H. (2001) Comparison of models predicting invertebrate assemblages for biomonitoring in the Fraser River catchment, British Columbia. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 1395–1410.

Ricklefs R.E. & Schluter D. (1993). *Species diversity in ecological communities: historical and geographical perspectives*. The University of Chicago Press, Chicago.

Rosenberg D.M. & Resh V.H. (1993) *Freshwater Biomonitoring and Benthic Macroinvertebrates*. Chapman and Hall, London.

Sokal R.R. & Rohlf F.J. (1987) *Biometry: the Principles and Practice of Statistics in Biological Research*. W. H. Freeman and Company, San Francisco.

Sovell L.A. & Vondracek B. (1999) Evaluation of the fixed-count method for Rapid Bioassessment Protocol III with benthic macroinvertebrate metrics. *Journal of the North American Benthological Society*, **18**, 420–426.

Stark J.D. (1993). Performance of macroinvertebrate community index: effects of sampling methods, sample replication, water depth, current velocity, and substratum on index values. *New Zealand Journal of Marine and Freshwater research*, **27**, 463–478.

Statzner B., Resh V.H. & Roux A. (1994). The synthesis of long-term ecological research in the context of concurrently developed ecological theory: design of a research strategy for the Upper Rhone River and its floodplain. *Freshwater Biology*, **31**, 253–263.

Stevens D. (1994). Implementation of a National Monitoring Program. *Journal of Environmental Management*, **42**, 1–29.

Stevens D.L. & Olsen A.R. (1999) Spatially restricted surveys over time for aquatic resources. *Journal of Agricultural, Biological, and Environmental Statistics*, **4**, 415–428.

Turner A.M. & Trexler J.C. (1997) Sampling aquatic invertebrate from marshes: evaluating the options. *Journal of North American Benthological Society*, **16**, 694–709.

Unicomarine Ltd. (1996). *National Marine Biological Analytical Quality Control Scheme*. Final Report Year 2, April 1995–March 1996, 26p. Unicomarine Ltd., England.

Wagner H.H. & Wildi O. (2002) Realistic simulation of the effects of abundance distribution and spatial heterogeneity on non-parametric estimators of species richness. *Ecosience*, **9**, 241–250.

Williams B.K., Nichols J.D. & Conroy M.J. (2001) *Analysis and Management of Animal Populations: Modeling, Estimation, and Decision Making*. Academic Press, New York.

Wright J. (1995) Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. *Australian Journal of Ecology*, **20**, 181–197.

Wright J.F., Sutcliffe D.W. & Furse M.T. (2000) *Assessing the Biological Quality of Fresh Waters: RIVPACS and Other Techniques*. Freshwater Biological Association, Ambleside, Cumbria, U.K.

Zar J.H. (1999) *Biostatistical Analysis*, 4th edn. Prentice Hall, New Jersey.

(Manuscript accepted 25 June 2003)